A Comprehensive Review of Multimedia Steganalysis Using Hybrid CNN–Transformer Deep Learning Models

1st Aroob Mukhtar
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
aroobmukhtar850@gmail.com

2nd M.Madni
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
madniakram246@gmail.com

3rd Umar Daraz
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
ru4004421@gmail.com

4th Farhan Hassan
Dept. of Information and Communication Engineering
The Islamia University of Bahawalpur
Bahawalpur, Pakistan
farhan.hassan@iub.edu.pk

***Abstract*—Multimedia steganography is the method of hiding information within images, audio, video, and text. The ability to detect hidden data in multimedia steganography is important for cybersecurity and secure communication systems. Historically, steganalysis techniques were conducted manually and proved to be less effective against sophisticated hiding techniques currently developed using deep neural networks and their hybrid models. Recent improvements in deep Convolutional neural networks and transformers have led to better detection of local and global patterns. In this review, we provide a comprehensive survey of recent developments in advanced multimedia steganalysis, particularly highlighting the hybrid approaches combining CNNs and Transformers, as well as CNNs and LSTMs. This review classifies the current steganalysis techniques based on existing studies, evaluates their performance on widely used benchmarks, examines architectural designs, and summarizes the latest advancements related to image, audio, video, and text-based steganalysis. This review highlights the need for new approaches like cross-media generalized steganalysis, multi-scale fusion methods, attention models, and LLM-based semantic steganalysis.**

***Index Terms*—Steganalysis, Steganography, Convolutional Neural Networks, Transformers, Deep Learning, Multimedia Security, Hybrid Models.**

I. Introduction

Digital communication has significantly transformed modern society. The demand for robust and secure communication systems is increasing to ensure information security. Although numerous secure methods have been proposed and implemented, ongoing development aims to enhance their effectiveness in both security and performance [1]. Steganography is a technique for the concealed transfer of information and can be described as the study of hidden communication, typically focusing on methods to obscure the presence of the transmitted message [2], [3]. Unlike cryptography, which aims to protect communications by transforming the data into a secure format for protection, steganography techniques are designed to conceal the message, making it challenging for an observer to determine where the message is located [4]. In text steganography, data is concealed through methods like altering word or line positions, making formatting adjustments, or replacing words with synonyms [5]. Meanwhile, image steganography incorporates information into either the spatial or frequency domain using techniques like least significant bit (LSB), discrete cosine transform (DCT), or discrete wavelet transform (DWT) [6]. Additionally, audio and video steganography embed messages within sound files or video sequences, employing methods such as parity coding, phase coding, spread spectrum, LSB, or DCT, which allows for the secure transfer of substantial data. Lastly, network protocol steganography takes advantage of communication protocols, such as TCP/IP or peer-to-peer services, to establish hidden channels for the exchange of secret messages [7], [8].

Steganalysis has been thoroughly examined over the past ten years. Its primary goal is to identify the existence of concealed messages within digital covers, such as images from a recognizable source. Typically, this challenge is presented as a binary classification issue to differentiate between cover and stego items [9], [10]. Steganalysis can be done on different types of cover media, such as images, audio, video, text, and network protocols. Its goal is to find hidden information that could be used for harmful purposes [11]. In image steganalysis, the focus can be on two domains: spatial and frequency [12]–[14], while video and audio steganalysis employ signals and frame-level patterns to determine whether the content is cover or stego [15], [16]. Network protocols and text analysis often investigate covert channels at the syntactic, semantic, or protocol level to safeguard against steganography [17], [18]. Steganalysis plays a crucial role as

DOI: https://doi.org/10.67535/tsp.000002.022
Published under the ISAISS 2026 Proceedings

it helps uncover concealed information hidden within digital assets, which could potentially be exploited for terrorist acts, cybercrime, data theft, and clandestine communication [19], [20]. Deep learning tackles the challenges that traditional steganalysis methods fail to address by automatically learning relevant features. Convolutional Neural Networks (CNNs), Transformers, and other hybrid models are now essential in identifying steganography in digital assets [21].

This paper reviews recent research on hybrid deep learning models for multimedia steganalysis. It presents a clear taxonomy of hybrid architectures, compares their performance on various media types, points out current research challenges, and suggests a unified framework for future work. The goal is to give researchers a thorough overview of hybrid CNN–Transformer methods and their promise for strong, cross-media steganalysis.

II. Review Methodology

This review uses a systematic process to select literature on recent advances in multimedia steganalysis. Mostly studies from 2019 to 2025 were gathered and a few other year papers also used from major digital libraries such as IEEE Xplore, Springer, Elsevier, arXiv and other digital repositories. The review focuses on peer-reviewed articles and high-quality preprints that discuss steganalysis with deep learning methods. The inclusion criteria targeted works related to image, audio, video, and text steganalysis that employed CNNs, Transformers, or hybrid architectures and provided architectural or performance-based analysis. Studies focusing solely on steganography, traditional machine learning approaches, or non-multimedia data were excluded. This methodology ensures balanced coverage of state-of-the-art techniques while maintaining relevance to contemporary multimedia steganalysis research.

III. Background

Steganalysis serves as the counterpart to steganography, concentrating on identifying and retrieving hidden text from the medium [22]–[24].

A. Traditional Methods

1) Feature Engineering: In machine learning, feature engineering is the process of selecting and extracting features from a dataset that help differentiate between stego and cover images [25]. Limitations: Feature engineering encounters several challenges, requiring domain expertise, and are not well generalized, along with issues related to scalability [26], [27].

2) Ensemble classifiers: An ensemble classifier combines various classifiers, including the weakest one, to create one robust model for supervised learning. The performance of this classifier model relies on the training of its components and the approach used for aggregating their votes [28]. Ensemble classifiers exhibit better scalability concerning the number of training samples and feature dimensions, achieving performance that rivals that of more intricate SVMs [29]. Limitations: Nonetheless, ensemble classifiers enhance accuracy however, their performance is dependent on feature quality and also face challenges in capturing complex patterns with new embedding techniques [29], [30].

3) Spatial Rich Models: The Spatial Rich Model (SRM) is a highly efficient technique for steganalysis. It leverages the statistics of adjacent noise residual samples as features to identify the changes in dependencies that result from embedding [31]. Limitations: Spatial Rich Models (SRM) experience significant computational demands and rely heavily on SVM and a lack of robustness against multimedia steganogrpahy [32].

B. Deep Learning Methods

Deep learning has addressed problems such as automatic detection of subtle embedding noise in images, videos, audio, text, and multimodal content has been simplified through the use of deep learning approaches such as CNN, Generative Adversarial Networks (GANs), Transformers, and hybrid architectures [33].

CNN and transformers enhanced modern multimedia steganalysis by using CNN to capture spatial features and local residual characteristics, while applying transformers to analyze global patterns and long-range dependencies across audio, video, and images [34].

IV. CNN Based Methods

A. Audio Steganalysis

Lin, Yuzhen, and colleagues present an innovative convolutional neural network (CNN) that improves audio steganalysis through the implementation of a uniquely designed convolutional layer, the use of truncated linear unit activation, and parameter transfer to identify subtle embedding noise [35].

Ren, Yanzhen, et al. develop a steganalysis technique MultispecNet that employs a multiscale spectrogram alongside a deep ResNet, a variety of CNN, to identify steganography in MP3 and AAC codecs by examining alterations in the time-frequency domain. This approach achieves greater accuracy compared to conventional techniques [36].

B. Video Steganalysis

Li, Zhonghao, et al. presents a PU-based Wide Residual-Net (PWRN) steganography method for HEVC videos that offers high embedding efficiency without sacrificing visual quality. The approach uses a super-resolution CNN with wide residual filters to reconstruct I-pictures, which helps improve P-picture prediction, lower the bitrate, and resist PU-targeted steganalysis [37].

SSTNet is a framework for detecting face manipulation. It analyzes spatial, steganalysis, and temporal features to find both visible and hidden signs of tampering [38], [39].

C. Image Steganalysis

Reinel, Tabares-Soto, et al. introduces a new CNN for spatial image steganalysis. It uses a preprocessing stage with filter banks, depthwise and separable convolutions, and skip connections to better highlight steganographic noise [40].

D. Text Steganalysis

Wen, Juan, et al. utilize a word embedding layer to capture the semantic and syntactic characteristics of words. Next, various rectangular convolution kernels are employed to identify features within sentences. To enhance performance further, we introduce a decision-making strategy aimed at recognizing lengthy texts [41].

V. Transformers-Based Methods

A. Audio Steganalysis

C. Zhang et al. introduce a model called TENet, a Transformer-based steganalysis framework for VoIP that utilizes codeword and position embeddings to uncover hidden representations of VoIP streams aimed at identifying QIM-based steganography [42].

B. Text Steganalysis

Wanli Peng et al. introduced a method for text steganalysis using transformers that leverages BERT to extract semantic features [43].

MB.Yang et al. investigate the application of large language models in the realm of linguistic steganalysis, leveraging the human-like characteristics of LLMs in text analysis to uncover concealed steganographic messages [44].

C. Image Steganalysis

In image steganalysis, transformers are not the sole architecture employed they are frequently integrated into hybrid models such as CNN-Transformers.

VI. Hybrid Architectures

A. Hybrid CNN-Transformers

1) Audio Steganalysis: A hybrid of CNN and Transformer, proposed by J. Peng et al., was designed for audio, which involves the extraction of features using multi-scale CNN and Transformer components [45].

The Dual-View VoIP Steganalysis Framework (DVSF) proposed by P. Zhou et al. utilizes a hybrid CNN-Transformer approach that simultaneously identifies detailed local characteristics and overarching streaming features to uncover steganography even at low embedding rates and brief duration circumstances [46].

2) Image Steganalysis: G.Lou et al. introduce a Convolutional Vision transformer for image steganalysis, which effectively captures both local and global relationships among noise features [47]. Mario Alejandro Bravo-Ortiz and colleagues present CVTStego-Net, a convolutional vision transformer designed for spatial domain image steganalysis, which combines the benefits of convolutional layers with the effectiveness of attention mechanisms to effectively capture both local and global relationships [48].

K. Wei et al. introduce a new steganalysis network that can detect stego images of any size without needing retraining. The network uses pixel difference convolution to extract local artifacts and an improved transformer module to capture global features [49].

Z. Zhou and colleagues introduce GTSCT-Net, a neural network designed to identify steganography in medical images by integrating global texture feature extraction with multi-head self-attention and deep convolution blocks [50].

W. Jiahao et al. introduce a hybrid model called CTS-Net (CNN-Transformer image steganography network) designed for image steganalysis. This model effectively captures dependencies in both local and global features of steganographic signals [51].

A convolutional transformer network has been introduced by KK. Wei et al. for the purpose of color image steganalysis. It starts with the preprocessing module, which produces residual images for each color channel. The residual images are truncated and concatenated.The residuals are processed in a hybrid module consisting of a CNN layer and a Transformer, which helps to incorporate local as well as global features pertaining to steganalytic images. This, in turn, relies upon global covariance pooling layers and two fully connected layers with a dropout technique [52].

B. Hybrid CNN-LSTM

1) Audio Steganalysis: Songbin Li and colleagues introduce a universal steganalysis technique that incorporates code element (CE) embedding, Bi-LSTM, and CNN with attention mechanisms [53]. For detecting voice signals, K.Li et al. present a steganalysis method that integrates 3D convolution with Bi-LSTM, alongside an attention mechanism. Initially, the encoded audio data is divided into three-dimensional data blocks, and 3D convolution is applied to capture the relations between successive frames. Following this, Bi-LSTM is utilized to derive the contextual features. Ultimately, the extracted features are merged and input into a classification network to ascertain whether hidden information is present in the speech stream [54].

E-SWAN is a hybrid CNN-LSTM model for real-time VoIP steganalysis, introduced by K. wang et al. It effectively captures both temporal and spatial features of speech. Experiments show that it performs better than existing methods when embedding rates are low and speech durations are short [55].

2) Image Steganalysis: Y. Shelar et al. propose a CNN-LSTM and an improved VGG-16 network for detecting image forgeries, including copy-move alterations. They use deep learning-based intra-frame forensic analysis. This approach overcomes the limitations of traditional methods and improves the detection of fake images, even when classifying datasets is difficult [56]. Shehab et al. present a hybrid CNN-LSTM model for image steganalysis. This model improves feature extraction by replacing the fully connected layers of CNN with LSTM [57].

VII. Comparison Tables

A. Image Steganalysis Table

The development of deep learning for image steganalysis is shown in Table 1. Several pioneering CNN-based models, such as Reinel et al. (2021), depend on residual learning and filter banks to model steganographic noise. Hybrid architectures

TABLE I
COMPARISON OF IMAGE STEGANALYSIS METHODS

Paper	Model	Dataset	Key Contribution
Reinel et al. (2021) [40]	GBRAS-Net (CNN)	BOSS, BOWS2	Residual-focused CNN using filter banks + separable convolutions.
Lou et al. (2022) [47]	CNN + ViT	BOSS, ALASKA2	Conv–ViT hybrid capturing local and global noise features.
Bravo-Ortiz et al. (2024) [48]	CVTStego-Net	ALASKA2	Three-stage Conv–ViT framework for spatial steganalysis.
K. Wei et al. (2024) [49]	PixelDiff CNN + Trans.	ALASKA2	Pixel-diff features + enhanced Transformer with GCP/SPP.
Z. Zhou et al. (2024) [50]	GTSCT-Net	Medical	Texture-aware Conv–Transformer improving medical detection.
W. Jiahao et al. (2025) [51]	CTS-Net	Large image sets	Multi-scale residuals + CNN–Transformer fusion.
KK. Wei et al. (2025) [52]	CTNet	Color images	Channel-wise residuals + CNN–Transformer feature groups.
Shelar et al. (2023) [56]	CNN–LSTM, VGG16	Forgery sets	Detects image forgeries and copy-move edits.
Shehab et al. (2025) [57]	CNN–LSTM	BOSS, BOWS, ALASKA2	LSTM-enhanced CNN for secure medical image analysis.

from Lou et al. (2022) and Bravo-Ortiz et al. (2024) combine CNN and Vision Transformers to capture both local and global features of images. Recent studies that have combined CNN with either Transformers or LSTMs include work by K. Wei et al., W. Jiahao et al., Shelar et al., and Shehab et al. These combinations help achieve high accuracy in detecting steganography across various image datasets, including BOSS, ALASKA2, and medical images.

B. Video Steganalysis Table

TABLE II
COMPARISON OF VIDEO STEGANALYSIS METHODS

Paper	Model	Dataset	Key Contribution
Li et al. (2022) [37]	PWRN (CNN)	HEVC	Wide residual CNN improving I/P-frame prediction and resisting PU attacks.
SSTNet (2020) [38], [39]	CNN + RNN	Face Forensics++	Detects manipulated faces via spatial, noise, and temporal cues.
Bouzegza et al. (2022)	Survey	Multiple	Comprehensive review of video steganalysis research.

Table II shows methods from surveys (Bouzegza et al., 2022), CNN/RNN models (SSTNet, 2020), and residual-based CNNs for HEVC videos (Li et al., 2022). It emphasizes the use of spatial and temporal features to detect manipulations.

C. Audio Steganalysis Table

Table III summarizes deep learning approaches for audio and speech steganalysis. Lin et al. (2019) improve CNN for detecting weak noises. Ren et al. (2022) use DeepResNet with a multiscale spectrogram for MP3 and AAC streams. Transformer-based models like TENet (2024) effectively encode VoIP streams. Hybrid CNN-Transformer architectures (Peng, SAT 2023; Zhou, DVSF 2025) capture both local and global features. Li et al. (2021), K. Li et al. (2025), and E-SWAN (2025) use CNN-LSTM and its variant, Bi-LSTM, for real-time and multi-method detection. This shows the benefits of combining temporal and spectral modeling.

TABLE III
COMPARISON OF AUDIO STEGANALYSIS METHODS

Paper	Model	Dataset	Key Contribution
Lin et al. (2019) [35]	CNN	Private	Special CNN layer + ReLU variant + parameter transfer for subtle noise detection.
Ren et al. (2022) [36]	Deep ResNet	MP3, AAC	Multiscale spectrogram + ResNet for time–frequency steganalysis.
Zhang et al. (TENet, 2024) [42]	Transformer	VoIP	Transformer encoder with codeword/position embeddings.
Peng et al. (SAT, 2023) [45]	CNN + Transformer	VoIP	Multi-scale CNN + improved Transformer for global/local features.
Zhou et al. (DVSF, 2025) [46]	Hybrid CNN–Transformer	VoIP	Dual-view detection for low-rate and short-duration VoIP steganography.
Li et al. (2021) [53]	CNN + Bi-LSTM	Compressed	CE embedding + Bi-LSTM + CNN for multi-method detection.
K. Li et al. (2025) [54]	3D CNN + Bi-LSTM	Custom	Extracts frame relations + contextual features.
E-SWAN (2025) [55]	CNN + LSTM	PMS, CNV	Real-time VoIP detection using sliding-window CNN–LSTM.

D. Text Steganalysis Table

TABLE IV
COMPARISON OF TEXT STEGANALYSIS METHODS

Paper	Model	Dataset	Key Contribution
Wen et al. (2019) [41]	CNN	Custom	First CNN-based text steganalysis using word embeddings + multi-kernel convolutions.
Peng et al. (2023) [43]	Transformer + Dual Attention	Multiple	Combines semantic (BERT) + statistical features; strong blind and semi-blind detection.
Yang et al. (2024) [44]	LLM-based	Multiple	LLM-powered generative modeling identifies subtle semantic inconsistencies; domain-agnostic solution.

Table IV shows that the techniques of text steganalysis range from early CNN-based methods to hybrid and attention-based models. Wen et al. (2019) were the first to develop CNN-based text steganalysis. Peng et al. (2023) introduced hierarchical dual-attention networks, which capture semantic and structural patterns. Yang et al. (2024) employs large language models to recognize human-like semantic manipulations. These methods highlight the trend of merging different approaches in representation learning for better text steganalysis.

VIII. KEY CONTRIBUTIONS

A. Key Contributions of this review

Below are the key contributions of this review paper:

1) Cross-Domain Coverage: This paper offers a comprehensive look at steganalysis in images, audio, video,

and text. This is different from earlier reviews that concentrate on just one type of media.

2) Focus on Hybrid Deep Models: We present the first structured analysis of hybrid CNN, Transformer, and CNN-LSTM architectures. We show how these models work together to capture both local noise and broader contextual patterns.
3) Comprehensive Taxonomy: We introduce a clear and systematic taxonomy that categorizes steganalysis approaches. These include traditional feature-based models, pure CNN models, Transformer-based methods, and hybrid deep learning architectures.
4) Identification of Research Gaps: The review examines the limitations of models in terms of generalisation, dataset availability, robustness, and computational cost, underscoring the significant difficulties encountered by contemporary multimedia steganalysis systems.
5) Foundation for Future Research: This paper emphasizes the emerging trends in steganalysis, including multi-scale fusion, global and local attention mechanisms, real-time steganalysis, LLM-based linguistic detection, and the hybrid architecture of CNN-Transformer for multimedia steganalysis across various media types, which opens new avenues for future research.

B. Limitations of Existing Steganalysis Methods

Existing steganalysis methods have the following disadvantages:

- Poor Generalization Ability Across Datasets and Media.
- Low robustness against adaptive attackers and deep learning-based embedding.
- Sensitive to compression, noise, and usual distortions.
- High computational cost, thus not suitable for real-time processing.
- Media concentration, with undetected cross-media overlaps.
- Detection of low embedding rates, particularly in audio/VoIP.
- Problems of text steganography in the contemporary world, particularly with semantic/generative text.

IX. Conclusion

This review shows that multimedia steganalysis is moving away from traditional CNN-based methods toward hybrid models that use Transformers, recurrent networks, and residual learning. Recent research finds that simply increasing model depth or relying on mathematical improvements does not always lead to better generalization. Instead, newer approaches focus on using residual components, attention mechanisms, and modeling global dependencies in images, audio, video, and text. Hybrid models like ResNet-ViT for images, Transformer-based pixel difference methods, and audio models that combine CNNs with Transformers or LSTMs are now leading the field. However, despite these advancements, several challenges remain. Existing models face issues in cross-media generalization, detecting at low embedding rates, implementing real-time solutions, and ensuring resilience against adaptive deep steganography are crucial challenges. Additionally, the limited availability of extensive, well-balanced datasets hinders the scalability of models, while larger datasets can lead to challenges such as increased computational expenses, overfitting, and diminished generalization. Overall, the findings suggest that hybrid CNN-Transformer architectures demonstrate promising performance in modern multimedia steganalysis. Future studies should focus on developing lightweight, integrated multimodal systems that utilize attention-based global features along with LLM techniques to improve efficiency, robustness, and real-world applicability in detecting multimodal steganography.

References

[1] D. A. Shehab and M. J. Alhaddad, "Comprehensive survey of multimedia steganalysis: techniques, evaluations, and trends in future research," Symmetry, vol. 14, no. 1, art. no. 117, 2022. [Online]. Available: https://www.mdpi.com/2073-8994/14/1/117

[2] S. D. Dickman, "An overview of steganography," Department of Computer Science, James Madison University Infosec Techreport, vol. 2, pp. 305–315, 2007.

[3] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," IEEE J. Sel. Areas Commun., vol. 16, no. 4, pp. 474–481, 2002.

[4] N. Hamid, A. Yahya, R. B. Ahmad, and O. M. Al-Qershi, "Image steganography techniques: an overview," Int. J. Comput. Sci. Secur. (IJCSS), vol. 6, no. 3, pp. 168–187, 2012.

[5] R. B. Krishnan, P. K. Thandra, and M. S. Baba, "An overview of text steganography," in Proc. 2017 Fourth Int. Conf. Signal Process., Commun. and Networking (ICSCN), 2017, pp. 1–6.

[6] N. Subramanian, O. Elharrouss, S. Al-Maadeed, and A. Bouridane, "Image steganography: a review of the recent advances," IEEE Access, vol. 9, pp. 23409–23423, 2021.

[7] N. Kaur and S. Behal, "A survey on various types of steganography and analysis of hiding techniques," Int. J. Eng. Trends Technol., vol. 11, no. 8, pp. 388–392, 2014.

[8] H. Singh, "Analysis of different types of steganography," Int. J. Sci. Res. Sci., Eng. Technol., vol. 2, pp. 578–582, 2016.

[9] Y. Qian, J. Dong, W. Wang, and T. Tan, "Deep learning for steganalysis via convolutional neural networks," in Media Watermarking, Security, and Forensics 2015, vol. 9409, pp. 171–180, 2015.

[10] M. Chaumont, "Deep learning in steganography and steganalysis," in Digital Media Steganography, Elsevier, 2020, pp. 321–349.

[11] V. Leask, R. Cogranne, D. Borghys, and H. Bruyninckx, "UNCOVER: Development of an efficient steganalysis framework for uncovering hidden data in digital media," in Proc. 17th Int. Conf. Availability, Reliability and Security, 2022, pp. 1–8.

[12] T. Reinel, R. Ramos-Pollán, and G. Isaza, "Deep learning applied to steganalysis of digital images: a systematic review," IEEE Access, vol. 7, pp. 68970–68990, 2019.

[13] K. Karampidis, E. Kavallieratou, and G. Papadourakis, "A review of image steganalysis techniques for digital forensics," *J. Inf. Secur. Appl.*, vol. 40, pp. 217–235, 2018.

[14] S. Wu, S. Zhong, and Y. Liu, "Deep residual learning for image steganalysis," *Multimedia Tools Appl.*, vol. 77, no. 9, pp. 10437–10453, 2018.

[15] H. Ghasemzadeh and M. H. Kayvanrad, "Comprehensive review of audio steganalysis methods," *IET Signal Process.*, vol. 12, no. 6, pp. 673–687, 2018.

[16] J. Kunhoth, N. Subramanian, S. Al-Maadeed, and A. Bouridane, "Video steganography: recent advances and challenges," *Multimedia Tools Appl.*, vol. 82, no. 27, pp. 41943–41985, 2023.

[17] Y. Shen, L. Huang, X. Lu, and W. Yang, "A novel comprehensive steganalysis of TCP/IP covert channels based on protocol behaviors and support vector machine," *Security Commun. Netw.*, vol. 8, no. 7, pp. 1279–1290, 2015.

[18] M. Taleby Ahvanooey, Q. Li, J. Hou, A. R. Rajput, and Y. Chen, "Modern text hiding, text steganalysis, and applications: A comparative analysis," *Entropy*, vol. 21, no. 4, p. 355, 2019.

[19] N. Provos and P. Honeyman, "Hide and seek: An introduction to steganography," *IEEE Security & Privacy*, vol. 1, pp. 32–44, Jun. 2003, doi: 10.1109/MSECP.2003.1203220.

[20] M. Dalal and M. Juneja, "Steganography and steganalysis (in digital forensics): a cybersecurity guide," *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5723–5771, 2021.

[21] N. J. De La Croix, T. Ahmad, and F. Han, "Comprehensive survey on image steganalysis using deep learning," *Array*, vol. 22, p. 100353, 2024.

[22] R. Cogranne and J. Fridrich, "Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2627–2642, 2015.

[23] T. Reinel, R. Ramos-Pollan, and G. Isaza, "Deep learning applied to steganalysis of digital images: A systematic review," *IEEE Access*, vol. 7, pp. 68970–68990, 2019.

[24] M. Chaumont, "Deep learning in steganography and steganalysis from 2015 to 2018," *arXiv preprint arXiv:1904.01444*, 2019.

[25] M. Anitha and M. Azhagiri, "Feature engineering techniques for Stegware analysis: An extensive survey," in *Proc. Int. Conf. Intell. Comput. Sustain. Dev.*, pp. 162–174, 2023.

[26] J. B. Guttikonda and S. R. Sridevi, "A new steganalysis approach with efficient feature selection and classification algorithms for identifying stego images," *Multimedia Tools Appl.*, vol. 78, no. 15, pp. 21113–21131, 2019.

[27] M. Boroumand and J. Fridrich, "Applications of explicit non-linear feature maps in steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 4, pp. 823–833, 2017.

[28] H. Bonab and F. Can, "Less is more: A comprehensive framework for the number of components of ensemble classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2735–2745, 2019.

[29] A. Torkaman and R. Safabakhsh, "A fast and accurate steganalysis using ensemble classifiers," in *Proc. 8th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, pp. 22–26, 2013.

[30] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 432–444, 2011.

[31] P. Wang, Z. Wei, and L. Xiao, "Pure spatial rich model features for digital image steganalysis," *Multimedia Tools Appl.*, vol. 75, no. 5, pp. 2897–2912, 2016.

[32] M. Goljan, J. Fridrich, and R. Cogranne, "Rich model for steganalysis of color images," in *2014 IEEE Int. Workshop Inf. Forensics Security (WIFS)*, pp. 185–190, 2014.

[33] M. Bouzegza, A. Belatreche, A. Bouridane, and M. Tounsi, "A comprehensive review of video steganalysis," *IET Image Process.*, vol. 16, no. 13, pp. 3407–3425, 2022.

[34] J. Wang, H. Yan, and J. Gu, "Image steganalysis based on CNN-Transformer," *J. Shenzhen Univ. Sci. Eng.*, vol. 42, no. 2, 2025.

[35] Y. Lin, R. Wang, D. Yan, L. Dong, and X. Zhang, "Audio steganalysis with improved convolutional neural network," in *Proc. ACM Workshop Inf. Hiding Multimedia Security*, pp. 210–215, 2019.

[36] Y. Ren, D. Liu, C. Liu, Q. Xiong, J. Fu, and L. Wang, "A universal audio steganalysis scheme based on multiscale spectrograms and DeepResNet," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 1, pp. 665–679, 2022.

[37] Z. Li, X. Jiang, Y. Dong, L. Meng, and T. Sun, "An anti-steganalysis HEVC video steganography with high performance based on CNN and PU partition modes," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 1, pp. 606–619, 2022.

[38] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *Proc. ICASSP*, 2020, pp. 2952–2956.

[39] M. Bouzegza, A. Belatreche, A. Bouridane, and M. Tounsi, "A comprehensive review of video steganalysis," *IET Image Process.*, vol. 16, no. 13, pp. 3407–3425, 2022.

[40] T. Reinel, H. B. Arteaga-Arteaga, M. A. Bravo-Ortiz, A. Mora-Rubio, D. Arias-Garzón, J. A. Alzate-Grisales, A. Burbano-Jacome, S. Orozco-Arias, G. Isaza, and R. Ramos-Pollan, "GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis," *IEEE Access*, vol. 9, pp. 14340–14350, 2021.

[41] J. Wen, X. Zhou, P. Zhong, and Y. Xue, "Convolutional neural network based text steganalysis," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 460–464, 2019.

[42] C. Zhang, S. Jiang, and Z. Chen, "TENet: Leveraging transformer encoders for steganalysis of QIM steganography in VoIP speech streams," *Multimedia Tools Appl.*, vol. 83, no. 19, pp. 57107–57138, 2024.

[43] W. Peng, S. Li, Z. Qian, and X. Zhang, "Text steganalysis based on hierarchical supervised learning and dual attention mechanism," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3513–3526, 2023.

[44] M. B. Yang, K. Pang, H. Wang, Y. Huang, *et al.*, "Towards next-generation steganalysis: LLMs unleash the power of detecting steganography," *arXiv preprint arXiv:2405.09090*, 2024.

[45] J. Peng, Y. Liao, and S. Tang, "VoIP steganalysis using shallow multi-scale convolution and transformer," in *Proc. Int. Conf. Security Privacy New Comput. Environ.*, pp. 326–351, 2023.

[46] P. Zhou, Z. Fang, Z. Yang, Z. Zhou, and L. Zhou, "Efficient streaming voice steganalysis in challenging detection scenarios," *IEEE Trans. Inf. Forensics Security*, 2025.

[47] G. Luo, P. Wei, S. Zhu, X. Zhang, Z. Qian, and S. Li, "Image steganalysis with convolutional vision transformer," in *ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 3089–3093, 2022.

[48] M. A. Bravo-Ortiz, E. Mercado-Ruiz, J. P. Villa-Pulgarin, C. A. Hormaza-Cardona, S. Quiñones-Arredondo, H. B. Arteaga-Arteaga, S. Orozco-Arias, O. Cardona-Morales, and R. Tabares-Soto, "CVTStego-Net: A convolutional vision transformer architecture for spatial image steganalysis," *J. Inf. Secur. Appl.*, vol. 81, p. 103695, 2024.

[49] K. Wei, W. Luo, and J. Huang, "Color image steganalysis based on pixel difference convolution and enhanced transformer with selective pooling," *IEEE Trans. Inf. Forensics Security*, 2024.

[50] Z. Zhou, K. Chen, D. Hu, H. Shu, G. Coatrieux, J. L. Coatrieux, and Y. Chen, "Global texture sensitive convolutional transformer for medical image steganalysis," *Multimedia Syst.*, vol. 30, no. 3, p. 155, 2024.

[51] J. Wang, H. Yan, and J. Gu, "Image steganalysis based on CNN-Transformer," *J. Shenzhen Univ. Sci. Eng.*, vol. 42, no. 2, 2025.

[52] K.-K. Wei, W.-Q. Luo, S.-Q. Tan, and J.-W. Huang, "CTNet: A convolutional transformer network for color image steganalysis," *J. Comput. Sci. Technol.*, vol. 40, no. 2, pp. 413–427, 2025.

[53] S. Li, J. Wang, P. Liu, M. Wei, and Q. Yan, "Detection of multiple steganography methods in compressed speech based on code element embedding, Bi-LSTM and CNN with attention mechanisms," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1556–1569, 2021.

[54] K. Li, F. Gao, and J. Yang, "Compressed speech steganalysis through deep feature extraction using 3D convolution and Bi-LSTM," *IEEE Access*, 2025.

[55] K. Wang, F. Gao, J. Yang, and H. Zhang, "E-SWAN: Efficient Sliding Window Analysis Network for Real-Time Speech Steganography Detection," *Computers, Materials & Continua*, vol. 82, no. 3, 2025.

[56] Y. Shelar, P. Sharma, and C. S. D. Rawat, "An Improved VGG16 and CNN-LSTM Deep Learning Model for Image Forgery Detection," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, pp. 73–80, 2023.

[57] D. Shehab and M. Alhaddad, "Image steganalysis using LSTM fused convolutional neural networks for secure telemedicine," *Frontiers in Medicine*, vol. 12, p. 1619706, 2025.